RC 22074 (99005) 24 May 2001 Chemistry/Statistics

Research Report

Scaling Invariant Principal Component Subset Selection in Principal Component Regression

Daniel E. Platt, Laxmi Parida, Yuan Gao, Isidore Rigoutsos
IBM T. J. Watson Research Center
P. O. Box 218
Yorktown Heights, NY 10598

Research Division
Almaden - Austin - Beijing - Haifa - India - T. J. Watson - Tokyo - Zurich

Scaling invariant principal component subset selection in principal component regression

Daniel E. Platt,* Laxmi Parida, Yuan Gao, Isidore Rigoutsos
IBM Thomas J. Watson Research Center, Yorktown Hgts, NY 10598

May 24, 2001

*rm 13-132F, IBM Thomas J. Watson Research Center,Yorktown Hgts, NY 10598, (914)945-1388, watplatt@us.ibm.com

Scaling invariant principal component subset selection in principal component regression

Abstract

Multiple regression has found applications in chemistry, pharmacology, and biochemistry as a tool for understanding molecular activity in quantitative structure activity relationship (QSAR) studies, in such diverse areas as near-infrared spectroscopy, mutational enzyme activity studies, and the analysis of gene expression data from chip arrays. Error analysis of principal component regression is dominated by the selection of an optimal subset of principal components, whose quality is measured by their contribution to the prediction of the independent variables and by their well conditioned behavior. Principal components are dependent on the scaling and units of measurement of the independent variables, which implies that the space spanned by some subset of principal components is not invariant to scaling transformations, yielding an arbitrary character. This paper presents a solution to the scaling problem in which a scale transformation is constructed which produces a set of equally well conditioned components, of which *one* contains all the predictive information of the regression. This scale transformation is independent of the initial scaling of the independent variables. This implies that the problems of conditioning and subset selection is an artifact of the initial scaling of the independent variables.

1 Introduction

Linear regression[1] has found application in chemistry, biology, and medicine, for of recognizing structural features important to the determination of chemical activity over a group of reactants in a poorly understood interaction. This is achieved by forming a linear relationship between variables (descriptors) describing the structural variation within the group of reactants, and the activities of the reactants. This relationship is called a quantitative structure activity relationship (QSAR). QSAR techniques have therefore found wide application in combinatorial chemistry studies in which variations in activities caused by the systematic modifications of structures can yield insight into the reaction activity mechanism.

A particularly interesting example of such a QSAR is comparative molecular field analysis (CoMFA),[2] in which reactants with common structural backbones varying by residue substitutions may be aligned with one another, and physical characteristics such as electrostatic potential and steric energies may be measured at each of thousands of oints on a common grid for each reactant. This is doubly interesting because it explicitly uses descriptors that describe the 3D character of a molecule rather than any description of its underlying topology, making it a pure 3D QSAR.

CoMFA QSAR regression systems are grossly underdetermined. Yet, it should be expected that the variations at the various field points should be correlated well enough (because of the relatively small number of residue substitutions as well as the discrete character of residue substitution in com-

binatoric studies) so that a meaningful relationship between those residue substitutions and the activities may be determined. Then those grid points around the residue sites that are important to the determination of the activity will make large contributions to the regression.

One procedure that treats such a grossly underdetermined system is the partial least squares (PLS) analysis,[3] which was popularized in the CoMFA program of TRIPOS's SYBYL package.[4] PLS has, since this popularization, emerged as a standard of analysis in numerous research publications.[5] The first applications of CoMFA have also emerged as benchmarks by which other 3D QSARs are measured.[2][6][7][8][9][10][11][12][13][14][15][16][17][18][19]

One aspect of electrostatic fields is that most of the volume of space surrounding a charge distribution is dominated by a relatively low number of parameters. The unique characterization of those parameters has led to the development of a 3D QSAR, called CoMMA, that does not depend on alignment of common backbones.[18][20] While loosing the detailed resolution of CoMFA, it has the advantage of codifying the longer-range characteristics of the molecular charge distribution in a small number of parameters and in a self-consistent manner.

The application of QSARs to spectral descriptors has a character similar to that of CoMFA. In this case, the spectra are sampled on a one-dimensional lattice of wavelengths. The amplitudes become structure variables for QSAR computations. The application of QSARs to spectroscopy, including near-infrared spectroscopy, shares the overdetermined character of CoMFA studies,[21] with a good review by Faber and Kowalski.[22]

QSARs have been applied to gene expression analysis in determining levels of gene expression as a function of descriptors of pharmacological substrates or treatment levels. [23] One particular application of regression of particular interest to gene expression studies in general involves the exploration of the relationship between transcriptional and translational control of gene expression.[24] One future application may be predicting survival of alleles, such as cancer survival rates [25], or perhaps of fermentation by yeast alleles [26] as a function of the levels of expression measured by gene array chips.

Yet another area of increasing interest is that of quantitative structure–property relationships (QSPRs). This is the prediction of physical characteristics such as boiling points, vapor pressure, critical temperature, critical micelle concentrations, polymer-glass transitions, for instance.[27] A possible candidate for application could be the phenomenological exploration of protein folding times. Enzyme mechanism is often elucidated by the examination of mutation activities, just as in combinatoric chemistry studies. Such studies have also been performed on protein folding rates.[28][29] This form of study is very consistent with standard combinatoric QSAR studies, and represents an opportunity for exploration by QSAR techniques.

Principal component regression (PCR) procedures [30][31] emerge naturally from the quadratic structure of the least-squares problem. The expression for the sum of the squares of the error between predicted and expected values may be expressed as a quadratic form in the regression coefficients. The principal components are orthogonal combinations of the data that di-

agonalize the coefficient quadratic form. Error propagation in PCR is simple and well understood.[30][31] Further, a least-squares χ^2 statistic providing a measure the goodness of fit based on a probability model is also commonly used. [31][32][33] The probability model in minimum χ^2 estimation assumes that the random deviations of the data from the linear regression model are Gaussian. Then the contributions of each component to the prediction of the dependent variables (defining their predictive power), as well as their contribution to the uncertainty in the regression coefficients (a measure of how well conditioned the component is). The problem of component subset selection has therefore become a dominating problem in studies of principal component linear regression and its application,[30][34][35] as well as other techniques of regression such as PLS,[3] where the quality of estimation as a function of the number of PLS components retained as well as the variation in regression coefficients is estimated via a cross-validation.[36] Other similar numerically intensive techniques include the bootstrap.[37]

This study emerged from consideration of the fact that the set of principal components are dependent on the scale and units of measurement of the various descriptor variables. This implies that component construction and subset selection is essentially arbitrary.[38] It was therefore desirable to try to construct an optimal scale matrix in which the selection of components would not be arbitrary, in which the conditioning would be much more uniformly well behaved, and in which a minimum number of components could be selected. The scaling transformation presented in this paper renders all components to be equally well conditioned, and permits the selection of a

single principal component which contains all of the predictive information.

Section 2 presents an analysis of the component subset selection problem, highlighting the problems of predictivity and well conditioning for review, and establishes notation. Section 3 presents a solution to the scaling problem and its implications to the number of predictive components as well as components that carry information about statistical uncertainty. Section 4 considers a simple application to the prediction of corticosteroid binding globulin activities[2] by CoMMA descriptors.[18] Section 5 presents conclusions.

2 Principal Component Subset Selection in Principal Component Regression

This section develops principal component regression, with particular consideration of the issues surrounding principal component subset selection.

The regression model predicts N sampled dependent variables y_i from independent N sampled D variables x_{ij} through model coefficients a_j, with uncontrolled variables accounting for a prediction error e_i, in the equation

$$y_i = \sum_j x_{ij} a_j + e_i. \tag{1}$$

Many regression studies use Greek letters to represent estimated regression parameters, but this usage is not universal.[31][33] The expected values of e_i are described as

$$E(e_i) = 0, \tag{2}$$

$$E(e_i e_j) = \Delta y_i^2 \delta_{ij}. \tag{3}$$

If each of the errors e_i is normally distributed, then the statistic

$$\mathcal{E}^2 = \sum_i \frac{e_i^2}{\Delta y_i^2} = \sum_i \frac{1}{\Delta y_i^2} \left(y_i - \sum_j x_{ij} a_j \right)^2 \tag{4}$$

is χ^2 distributed with N degrees of freedom. [30][31][32][33] This may be expressed in terms of matrices as

$$\mathcal{E}^2 = (Xa - y)^T C (Xa - y) \tag{5}$$

where C is the diagonal matrix with elements $(C)_{ij} = \delta_{ij}/\Delta y_i^2$, or more generally,

$$C^{-1} = \mathrm{cov}(y, y). \tag{6}$$

This may be expressed alternatively as

$$\mathcal{E}^2 = (a - a_0)^T X^T C X (a - a_0) + y^T C y - a_0^T X^T C X a_0, \tag{7}$$

where

$$a_0 = \lim_{\epsilon \to 0} (X^T C X + \epsilon I)^{-1} X^T C y. \tag{8}$$

The limit $\lim_{\epsilon \to 0} (X^T C X + \epsilon I)^{-1} X^T C^{1/2}$ is called a "generalized inverse" of $C^{1/2} X$. The limit $\lim_{\epsilon \to 0} (X^T C X + \epsilon I)^{-1}$ is undefined unless the matrix first operates on another matrix or vector which has no projections along eigenvectors of $X^T C X$ that correspond to eigenvalues equal to zero. However, if u is an eigenvector of $X^T C X$ with a zero eigenvalue $(Xu)^T C (Xu)$, it follows that any projection Xu of X along u will be zero since $X^T C X u = 0$ implies that $u^T X^T C X u = (Xu)^T C (Xu) = 0$. Since C is diagonal, this implies that each $(Xu)_i = 0$. Further, this implies that $u^T a_0 = 0$. Note that this solution

is not unique. Any $a_0' = a_0 + \delta a$, where $X^T C X \delta a = 0$, produces an equivalent $\mathcal{E}^2$.

Since $\mathcal{E}^2$ is a χ^2 statistic, it follows that a_0 and $X^T C X$ are essential statistics.[39] Any changes in a may be accounted for by the contribution to the error $\mathcal{E}^2$ by the coefficients

$$\mathcal{E}^2_{\text{coef}} = (a - a_0)^T X^T C X (a - a_0), \tag{9}$$

with the remainder accounted for by the residual

$$\mathcal{E}^2_{\text{res}} = y^T C y - a_0^T X^T C X a_0, \tag{10}$$

so that

$$\mathcal{E}^2 = \mathcal{E}^2_{\text{coef}} + \mathcal{E}^2_{\text{res}}. \tag{11}$$

The total number of degrees of freedom in $\mathcal{E}^2$ is N. The number of degrees of freedom in $\mathcal{E}^2_{\text{coef}}$ is equal to the number D_0 of eigenvectors with corresponding nonzero eigenvalues of $X^T C X$. This leaves $N - D_0$ degrees of freedom for $\mathcal{E}^2_{\text{res}}$. This partition is very reminiscent of Bayesian treatments of linear regression,[39] but the presentation here follows "frequentist" notions of sampling theory.

The expectation value of a_0 is

$$E(a) = a_0 = \lim_{\epsilon \to 0} (X^T C X + \epsilon I)^{-1} X^T C y. \tag{12}$$

The covariance is predicted by

$$\text{cov}(a, a) = E[(a - a_0)(a - a_0)^T] = \lim_{\epsilon \to 0} (X^T C X + \epsilon I)^{-1} \tag{13}$$

determined by the inverse of the quadratic coefficients in $\mathcal{E}^2_{\text{coef}}$. As pointed out before, this limit does not exist if there are eigenvalues of X^TCX equal to zero. This means that any contribution to a of any magnitude in a direction corresponding to a null eigenvector of X^TCX will not contribute anything to $\mathcal{E}^2$. This implies that the coefficients essentially have an infinite uncertainty and are completely undetermined in any underdetermined system. This is simply a reflection of the ambiguity in underdetermined systems.

A meaningful alternative measure of covariance is the amount by which the estimate of a will vary given the variations in y. This is essentially equivalent to the effect of allowing y to vary according to the variation in e. This implies

$$\begin{aligned}
\text{cov}_{\text{subspace}}(a,a) &= E\{a_0[e]a_0[e]^T\} \\
&= \lim_{\epsilon\to 0} E\{(X^TCX+\epsilon I)^{-1}X^TCee^TCX(X^TCX+\epsilon I)^{-1}\} \\
&= \lim_{\epsilon\to 0}(X^TCX+\epsilon I)^{-1}X^TCE(ee^T)CX(X^TCX+\epsilon I)^{-1} \\
&= \lim_{\epsilon\to 0}(X^TCX+\epsilon I)^{-1}X^TCC^{-1}CX(X^TCX+\epsilon I)^{-1},
\end{aligned}$$

or

$$\text{cov}_{\text{subspace}}(a,a) = \lim_{\epsilon\to 0}(X^TCX+\epsilon I)^{-1}X^TCX(X^TCX+\epsilon I)^{-1}. \qquad (14)$$

This limit does exist because

$$P_0 = \lim_{\epsilon\to 0}(X^TCX+\epsilon I)^{-1}X^TCX \qquad (15)$$

is a projection operator that picks out only those eigenvectors with nonzero eigenvalues. This expression compares favorably with the variation in the coefficients observed between the various regressions produced by cross-validation.

Such a result constitutes an explicit measure of the stability of the coefficients to variations in the dependant variables.

It is important to realize that while some consistency may be expected within a dataset, and it is possible to ask whether a model is consistent with a dataset in a statistical sense, that underdetermined systems do not yield definitive measurements of all the coefficients. Comparison with other datasets that could ultimately produce a complete model if the data were combined would not produce coefficients consistent with one another.

Not only is it possible for a regression to be underdetermined in the sense of having zero valued eigenvalues, but some of the eigenvalues of $X^T C X$ may be very small. Such a system is called "poorly conditioned." This corresponds to some $\mathrm{var}(a_i)$ being very large. Such terms can add spurious and large contributions to a_0 without significantly affecting $\mathcal{E}^2$. This suggests that it is desirable to exclude contributions from various subsets of components that may not correspond to zero valued eigenvalues.

The systematic consideration of the character of individual principal components in the analysis of the $\mathcal{E}^2$ quadratic form is perhaps the best definition for principal component regression. Consider a projection operator P that is a projection onto a subset K of eigenvectors u_k of $X^T C X$. As such, P satisfies

$$P = \sum_{k \in K} u_k u_k^T, \tag{16}$$

$$P^2 = P, \tag{17}$$

$$\left[P, X^T C X\right] = 0. \tag{18}$$

The effect of excluding components that project orthogonally to P is to require that any component of a projecting along

$$Q = P_0 - P \tag{19}$$

are zero so that

$$Qa = 0 \tag{20}$$

leaving

$$\mathcal{E}^2[P] = (Pa - a_0)^T X^T C X (Pa - a_0) + y^T C y - a_0^T X^T C X a_0.$$

It is desireable to repartition the contributions $\mathcal{E}^2_{\text{coef}}$ and $\mathcal{E}^2_{\text{res}}$ to reflect this projection:

$$\begin{aligned}(Pa - a_0)^T X^T C X (Pa - a_0) &= (Pa - Pa_0)^T X^T C X (Pa - Pa_0) \\ &+ (Qa_0)^T X^T C X (Qa_0),\end{aligned}$$

and that

$$(a_0)^T X^T C X (a_0) = (Pa_0)^T X^T C X (Pa_0) + (Qa_0)^T X^T C X (Qa_0),$$

it follows that

$$\mathcal{E}^2[P] = \mathcal{E}^2_{\text{coef}}[P] + \mathcal{E}^2_{\text{res}}[P], \tag{21}$$

where

$$\mathcal{E}^2_{\text{coef}}[P] = (Pa - Pa_0)^T P X^T C X P (Pa - Pa_0), \tag{22}$$

$$\mathcal{E}^2_{\text{res}}[P] = y^T C y - a_0^T P X^T C X P a_0. \tag{23}$$

This is a very suggestive partition of the degrees of freedom. The operator P removes degrees of freedom from $\mathcal{E}^2_{\text{coef}}[P]$ and essentially transfers them to $\mathcal{E}^2_{\text{res}}[P]$. Since the total number of degrees of freedom in $\mathcal{E}^2[P]$ remains N, and the number of degrees of freedom in $\mathcal{E}^2_{\text{coef}}[P]$ is now D_P, the total number of degrees of freedom in $\mathcal{E}^2_{\text{res}}[P]$ is now $N - D_P$. Further, while the degrees of freedom in $\mathcal{E}^2_{\text{res}}[P]$ increases when poorly conditioned eigenvectors are excluded, so does the total value of $\mathcal{E}^2_{\text{res}}[P]$. The relationship between the goodness of fit error and the exclusion of particular components is well understood. [30]

The partitioning of the error according to contributions by projections of components suggests an immediate application. It becomes possible to compare the goodness of fit for different subsets of components. In particular, for a subset of components projected by P, the probability that a χ^2 larger than this might be observed is $P(\chi^2_{N-D_P} > \mathcal{E}^2_{\text{res}}[P])$. Those with larger probabilities better represent the fit. Note that if $N = D_P$, which happens when all of the non-null components are used in an underdetermined system, then $P(\chi^2_{N-D_P} > \mathcal{E}^2_{\text{res}}[P])$ is undefined. There is essentially no statistical information about the quality of the fit if all of the principal components are included.

Further, the contributions of each individual component may also be determined. The contribution to $\mathcal{E}^2_{\text{res}}[P]$ may be determined for any particular component k. For any component k with eigenvector u_k, the projection operator is $P_k = u_k u_k^T$. This implies that the effect of any particular eigenvector

is to subtract a variation

$$\begin{aligned}\mathcal{E}_k^2 &= a_0^T u_k u_k^T X^T C X u_k u_k^T a_0 \\ &= y^T C X (X^T C X + \epsilon I)^{-1} u_k u_k^T X^T C X u_k u_k^T (X^T C X + \epsilon I)^{-1} X^T C y\end{aligned}$$

or

$$\mathcal{E}_k^2 = \frac{y^T C (X u_k)(X u_k)^T C y}{(X u_k)^T C (X u_k)}, \tag{24}$$

where $A_k = (X u_k)^T C (X u_k)$ is the eigenvalue of $X^T C X$ corresponding to eigenvector u_k. The contribution of the kth component to $\mathrm{cov}(a, a)$ varies as $1/A_k$. This is a reflection of how well conditioned the contribution is from this component. Small A_k components contain little discriminating information compared to the uncertainty they contribute to the regression coefficients. Exclusion of the smallest A_k contributions therefore improves the stability of the coefficients and reduces the size of the uncertainty in those parameters.

However, the largest $\mathcal{E}_k^2$ contribute the most towards improving the goodness of fit since they reduce $\mathcal{E}_{\mathrm{res}}^2[P]$ the most. Therefore, the value of $\mathcal{E}_k^2$ represents the predictive power of the kth component. It is possible therefore to rank the components by predictive power. Then, it is possible to construct a list of subsets with the largest predictive power, then the next list containing the largest together with the second largest, and then the third list containing the top three predictive components, etc. This reduces the computation from all 2^N possible subsets of components to a simple list of subsets N long. Once this is done, it is possible to compute $P(\chi^2_{N-D_P} > \mathcal{E}_{\mathrm{res}}^2[P])$ for each of the subsets. This probability generally goes through some extremum,

which represents the optimal subset of components. Since the questions of the information in a component, as measured by A_k and the contribution the component makes to the goodness of fit are distinct, exclusion of low information components may be achieved by applying a cutoff to A_k. A selection of the most important contributors to the goodness of fit may then be applied.

This approach has been inverted to consider the situation in which the size of the uncertainty is unknown, and it is desired to estimate some best uncertainty from the regression of the data. This may be achieved by choosing $C = I/\Delta Y^2$, to yield

$$E(\mathcal{E}^2_{\rm res}[P]) = N - D_P = \frac{1}{\Delta Y^2}\left(y^T y - a_0^T P X^T X P a_0\right), \qquad (25)$$

and solving for ΔY^2. The best subset is the one that produces the smallest ΔY^2. This component selection criterion is essentially identical to one proposed by Lott, [34], who also recognized the possibility of reducing the optimal space of subsets by ranking the components. However the connection between the selection of an optimal subset and a minimum ΔY^2 was not established, and connection with χ^2 was not explored. Generally, for overdetermined systems, ΔY goes through a minimum as the number of components is decreased. The smallest set is the best. However, in underdetermined systems there tends to be no minimum in ΔY. For a fixed ΔY, there will usually be some particular subset of components where $P(\chi^2_{N-D_P} > \mathcal{E}^2_{\rm res}[P])$ minimizes. Once some ΔY is selected and the component subset is extracted, the values of a_0 and $\mathrm{cov}(a, a)$ which are consistent with the quality of the

regression and the variation in the data may be computed.

The problem with the simple method of selecting a subset that spans the space of variation, as in PCA, is that the dependent variable may depend on some of the components with smaller variation. It is similar to trying to describe a pizza pie with a pancake model: if the short axis is discarded, there is no dimension to describe the layering of ingredients. This problem is well known, and there have been a number of solutions posed for selecting some optimal subset of components.[30] Yet many commercial packages still rank components according to their variation $A_k = (Xu_k)^T C(Xu_k)$, which only measures the range of variation of the component, and *not* the contribution of the component $\mathcal{E}_k$ to the goodness of fit.

3 Scale Invariance Problems in Component Selection

The previous section outlined the effect of principal component subset selection based on conditioning and predictive power. However, there is a major problem in the selection of principal components. This is that the components depend on the scaling, or units of measurement, of the independent variables. Simply changing the units of measurement can significantly affect the components, how well they are conditioned and how much predictive power they express. This section presents a simple scaling transformation that renders the system completely well conditioned, and which reduces the number of predictive components to one.

Define

$$\Gamma^2 = X^T C X. \tag{26}$$

Then

$$\lim_{\epsilon \to 0} (\Gamma + \epsilon I)^{-1} X^T C X (\Gamma + \epsilon I)^{-1} = P_0. \tag{27}$$

Define

$$\lim_{\epsilon \to 0} X (\Gamma + \epsilon I)^{-1} = V$$

Further define

$$\begin{aligned} \alpha &= \Gamma a, \\ \alpha_0 &= \Gamma a_0 = V^T C y, \end{aligned}$$

so that

$$X a = V \alpha. \tag{28}$$

In this representation, each component is equally well conditioned, and any orthogonal set of vectors spanning P_0 is a set of eigenvectors, and is thus a set of principal components.

The residual error is then

$$\begin{aligned} \mathcal{E}^2_{\text{res}} &= y^T C y - a_0^T X^T C X a_0 \\ &= y^T C y - \lim_{\epsilon \to 0} y^T C X \left(X^T C X + \epsilon I \right)^{-1} X^T C X \left(X^T C X + \epsilon I \right)^{-1} X^T C y \\ &= y^T C y - \lim_{\epsilon \to 0} y^T C V \Gamma \left(\Gamma^2 + \epsilon I \right)^{-1} \Gamma^2 \left(\Gamma^2 + \epsilon I \right)^{-1} \Gamma V^T C y \\ &= y^T C y - y^T C V P_0 V^T C y \\ &= y^T C y - \alpha_0^T P_0 \alpha_0. \end{aligned}$$

Again considering projection operators $P + Q = P_0$, where $PP_0 = P$, yields

$$\mathcal{E}^2_{\rm res} = y^T C y - \alpha_0^T P \alpha_0 - \alpha_0^T Q \alpha_0.$$

Now, define $P = \hat{\alpha}_0 \hat{\alpha}_0^T$ selects only one component $\hat{\alpha}_0$. The projection in Q is $\alpha_0^T Q \alpha_0 = 0$, leaving

$$\mathcal{E}^2_{\rm res}[P] = y^T C y - \alpha_0^T \alpha_0. \tag{29}$$

Under this scaling, constructed in a way that is independent of the descriptors' initial units of measurement, there are no poorly conditioned components, and the predictive information may be contained by only one component.

Another question that this scaling has bearing on is the distinction between the least error in the dependent variable y and the least error in the coefficients a.[40] Consider the transformation

$$Xa = \lim_{\epsilon \to 0} X(\Gamma + \epsilon I)^{-1} \Gamma a = V \alpha,$$

where

$$\alpha = \Gamma a. \tag{30}$$

Then the χ^2 error has the form

$$\begin{aligned} \mathcal{E}^2 &= y^T C y - a_0^T X^T C X a_0 + (a - a_0)^T X^T C X (a - a_0) \\ &= \lim_{\epsilon \to 0} \Big\{ y^T C y - \alpha_0^T (\Gamma + \epsilon I)^{-1} \Gamma^2 (\Gamma + \epsilon I)^{-1} \alpha_0 \\ &\quad + (\alpha - \alpha_0)^T (\Gamma + \epsilon I)^{-1} \Gamma^2 (\Gamma + \epsilon I)^{-1} (\alpha - \alpha_0) \Big\} \end{aligned}$$

or

$$\mathcal{E}^2 = y^T C y - \alpha_0^T P_0 \alpha_0 + (\alpha - \alpha_0)^T P_0 (\alpha - \alpha_0). \tag{31}$$

In this scaling, the measures of the errors in α are on the same basis as the measures in the errors in the prediction of the y.

Further, the covariance of α is

$$\begin{aligned} \mathrm{cov}(\alpha, \alpha) &= \Gamma \mathrm{cov}(a, a) \Gamma \\ &= \lim_{\epsilon \to 0} \Gamma (\Gamma^2 + \epsilon I)^{-1} \Gamma^2 (\Gamma^2 + \epsilon I)^{-1} \Gamma \\ &= P_0. \end{aligned}$$

So while only one component carries information about the regression, all of the components carry uncertainty. This implies that the covariance of a may be partitioned into the information bearing part $P = \hat{\alpha}_0 \hat{\alpha}_0^T$, and $Q = P_0 - P$, so that

$$\begin{aligned} \mathrm{cov}(a, a) &= \lim_{\epsilon \to 0} (\Gamma + \epsilon I)^{-1} \mathrm{cov}(\alpha, \alpha) (\Gamma + \epsilon I)^{-1} \\ &= \lim_{\epsilon \to 0} (\Gamma + \epsilon I)^{-1} P_0 (\Gamma + \epsilon I)^{-1} \\ &= \lim_{\epsilon \to 0} (\Gamma + \epsilon I)^{-1} (P + Q) (\Gamma + \epsilon I)^{-1} \\ &= \lim_{\epsilon \to 0} (\Gamma + \epsilon I)^{-1} P (\Gamma + \epsilon I)^{-1} + \lim_{\epsilon \to 0} (\Gamma + \epsilon I)^{-1} Q (\Gamma + \epsilon I)^{-1}. \end{aligned}$$

Each term is positive definite, making positive contributions to $\mathrm{var}(y)$ for some predicted y.

According to the notion of subset selection, the goal is to discard components that contain no predictive information, but which add uncertainty to the problem. In this case, the Q partition contains no predictive information, but does contain information about uncertainty. Essentially, the Q

contributions represents the uncertainty in estimating 0 given the quality of the data set.

A few points should be reviewed. First, for underdetermined systems, *any* contributions to a from $Q_0 = I - P_0$ produce no change in the error $\mathcal{E}^2_{\rm coef}$. Essentially, the a's are undetermined, or are only determined to within specified uncertainties within the subspace P_0. Within that subspace, the total uncertainty may be partitioned into contributions from the various components. While most of the components appear to contribute no information about value, they *do* represent regions of the space explored and spanned by variations in the y represented by $C^{-1} = {\rm cov}(y, y)$. So, besides transferring degrees of freedom from $\mathcal{E}^2_{\rm coef}$ to $\mathcal{E}^2_{\rm res}$, a subset selection P *also* projects out contributions to the uncertainty from the discarded components $Qa = 0$. As such, it no longer reflects the total space of a explored by the sampling in the X and ${\rm cov}(y, y)$.

Given the estimated parameters, it is possible to estimate a $y_{\rm est}$ (scalar) for a vector x of independent variables as

$$y_{\rm est} = x^T a_0 \pm \sqrt{x^T {\rm cov}(a, a) x}. \tag{32}$$

Following the previous section, consider the problem of estimating error bars from data. The first step is to identify $C^{-1} = \Delta Y^2 I$. Then

$$\begin{aligned} \mathcal{E}^2(\Delta Y) &= \frac{\mathcal{E}^2(1)}{\Delta Y^2} \\ \mathcal{E}^2_{\rm coef}(\Delta Y) &= \frac{\mathcal{E}^2_{\rm coef}(1)}{\Delta Y^2} \\ \mathcal{E}^2_{\rm res}(\Delta Y) &= \frac{\mathcal{E}^2_{\rm res}(1)}{\Delta Y^2} \end{aligned}$$

$$\begin{aligned}\mathcal{E}^2_{\text{coef}}[P](\Delta Y) &= \frac{\mathcal{E}^2_{\text{coef}}[P](1)}{\Delta Y^2} \\ \mathcal{E}^2_{\text{res}}[P](\Delta Y) &= \frac{\mathcal{E}^2_{\text{res}}[P](1)}{\Delta Y^2}.\end{aligned}$$

The expectation value $E\left[\mathcal{E}^2_{\text{coef}}[P](\Delta y)\right]$ should satisfy

$$\begin{aligned}E\left[\mathcal{E}^2_{\text{coef}}[P](\Delta Y)\right] &= E\left[\mathcal{E}^2[P](\Delta Y)\right] - E\left[\mathcal{E}^2_{\text{res}}[P](\Delta Y)\right] \\ &= N - D_P \\ &= \frac{E\left[\mathcal{E}^2_{\text{coef}}[P](1)\right]}{\Delta Y^2},\end{aligned}$$

where N is the number of data points, and D_P is the number of components spanned by P. This implies

$$\Delta Y^2 = \frac{E\left[\mathcal{E}^2_{\text{coef}}[P](1)\right]}{N - D_P}. \tag{33}$$

In the representation used in the previous section, it would appear that successive exclusion of principal components would cause ΔY^2 to pass through a minimum. However, in this representation, where all of the predictive information is contained in only *one* component, it is clear that there will be no minimum down to $D_P = 1$, and that a minimum where $D_P \neq 1$ is an artifact of the diagonalization of $X^T C X$ in whatever the independent variables' scalings were.

At the same time,

$$\text{cov}(a, a)(\Delta Y) = \text{cov}(a, a)(1)\Delta Y^2.$$

Discarding degrees of freedom, even though they do not predict y, will minimize D_P, increase $N - D_P$, reducing the expected value ΔY^2. Further, the

corresponding exclusion of those estimates in

$$\mathrm{cov}_P(a,a) = \lim_{\epsilon\to 0}(\Gamma+\epsilon I)^{-1}P(\Gamma+\epsilon I)^{-1}$$

further underestimates $\mathrm{cov}(a,a)$, which shows up in error bars $\Delta y_{\mathrm{est}} = \sqrt{x^T \mathrm{cov}(a,a)x}$ that are too small to be consistent with the regression data.

The problem of assessing the statistical quality of the data is particularly problematical for underdetermined systems. In this case, $\Delta Y = 0$, and there are no error bars.

4 QSAR Analysis of CoMMA: Methods and Results

The CoMMA descriptors[18] of a molecule are a vector of the form

$$I_x \quad I_y \quad I_z \quad p_x \quad p_y \quad p_z \quad p \quad q \quad q_{xx} \quad q_{yy} \quad d_x \quad d_y \quad d_z,$$

where I_x, I_y, and I_z are the principal values of the moment of inertia, p_x, p_y, p_z, are the absolute values of the components of the dipole moment written in the inertial principal axis frame, p is the magnitude of the dipole moment (invariant under translation in a neutral molecule). Since the quadrupole depends on the center of expansion if the molecule has a dipole, the field is not uniquely specified unless a unique center of expansion is specified. Given some expansion about an origin, translation of the center of expansion to the point specified by $\mathbf{Q}\cdot\vec{p} = \vec{0}$ yields a unique quadrupole moment. Further, since the dipole is an eigenvector of the quadrupole moment with eigenvalue 0, and the trace of the quadrupole $\mathrm{Tr}\mathbf{Q} = 0$, it follows there is

only one characteristic number q, along with the principal axes, that specifies the quadrupole.[20] While the unique specification of such a center is not absolutely necessary for a good QSAR, it does provide an index which may be used with a QSAR in conjunction with database searches on uniquely characterized molecular field characteristics. The numbers q_{xx} and q_{yy} represent the projections of the quadrupole on the dipolar axes. The vector d_x, d_y, and d_z is the absolute values of the displacement components from the center of mass to the center of dipole expressed in the inertial frame. Since the sense of the inertial axes is not determined, only the absolute values of the components of $\vec{p}$ and $\vec{d}$ were retained, and the quadrupolar cross terms q_{xy}, q_{xz}, and q_{yz} were discarded.[18][19] This implies the CoMMA descriptors are unchanged by chiral transformations, and prediction of chirally sensitive activites for chiral isomers will fail.

The data used for an illustrative example in this section have all been publish previously.[18][19] The 21 steroids distributed with the SYBYL[4] molecular modeling program are analyzed as an example. The twelfth steroid provided in that set, and listed as molecule 2 in the original citation,[2] has been recognized to be incorrect.[13][17] The structure used here was corrected. The 21 steroids are numbered and displayed in Figure 1. The structures were initially constructed with standard angles and bond lengths provided in the SYBYL 6.01 program,[4] global energy minimization based on the TRIPOS force field program,[41] with 10° angular resolution systematic search, followed by a further force-field optimization. The dipole and quadrupol moments of the charge distributions were computed from single

point Gaussian 92[42] computations on a 631G** basis set. The values are enumerated in Table 1. The activities to be predicted are listed in Table 2.

The computation outlined in Section 3 was performed, using all of the components in P_0 to predict the error. The independent variables were augmented with an auxilliary fourteenth "descriptor" whose value is always set equal to 1. This provides a simple way to include the offset in the error propagation of the coefficients. The predicted coefficients together with their expected errors are listed in Table 3. The size of the error bars would suggest that many of the coefficients do not contain sufficient signal to predict any meaningful regression. However, Table 4 shows very strong correlations between the coefficients that must be taken into account when computing a propagation of error. When this computation is properly performed, the resulting error bars are quite consistent with observation, as seen in Table 5, suggesting that some of the notions of controlling for poorly conditioned variables might be better managed through consideration of coefficient correlations.

The same computation was performed using the projection $P = \hat{\alpha}_0 \hat{\alpha}_0^T$, yielding unreasonably small error bars for the coefficients in Table 6, extremely strong correlations between the coefficients, Table 7, and error bars that do not reflect the observed deviations, as seen in Table 8. This ideal case of principal component subset selection, where the one component carrying predictive information was isolated and used, is seen not to contain sufficient information to correctly predict the error propagation.

5 Conclusions

Since there is a way of extracting a scaling transformation of the independent variables such that all the principal components are equally well conditioned, and in which only one component contains all the predictive information, it follows that the problem of principal component subset selection determined by conditioning and predictive contribution is an artifact of the initial scaling of the independent variables. Further, the distinction between the errors in the estimation of the coefficients as opposed to the estimation of the errors in the prediction of the dependent variable vanishes in the scaling proposed here. However, all of the information in this scaling is also contained in the components of the possibly more poorly conditioned initial scaling of the data.

If that scaling was poorly conditioned, then the coefficients will be highly correlated with large propagated errors, but those coefficients may produce good predictions of the activities. This further implies that slight variations in the data, or the inclusion or exclusion of data points, may produce wildly differing coefficients which nevertheless predict the same good estimated activities with reasonable uncertainties. If consistency in the predicted coefficients is important, then some component subset selection may be preferred. However, error propagation is invalidated, and such a selection must become an artifact of the accident of the initial selection of the physical units and the correlations in the sampled independent variables.

Another immediate conclusion is that it is not meaningful to infer the

importance of the various combinations of descriptors from the predictivity of various components, or to infer the predictive power of independent descriptors, since the predictive information can be contained by just one component while an effective description of the uncertainties requires all of the components.

An alternative approach might be to identify strongly correlated variables from the error analysis, and drop one of them from the regression. Such information may be more effectively obtained by using the correlations to determine which variables may be carrying redundant information that could confound the conditioning of the regression, and to use techniques such as cross-validation to identify a most effective subset of descriptors to describe the data.

In grossly overdetermined systems, such as CoMFA,[2] rapid and easily implemented algorithms such as PLS[3] have become the de facto standard. And, as has been seen earlier, it is very problematical to perform an error analysis for underdetermined systems except through cross-validation or bootstrapping techniques. Even though error analysis is difficult due to the nonlinearities in PLS, linearized approximations are available.[22] However, if the number of components is limited by a relatively small number of measurements, some common algorithms permit the rapid computation of leading principal components for systems with large numbers of independent descriptors.

Acknowlegement – The author wishes to thank B. David Silverman for many helpful discussions over many years on this topic, and for the use of

data he painstakingly collected.

References

[1] Pearson, K.; "On Lines and Planes of Closest Fit to Systems of Points in Space," *Philosophical Magazine*, (6) 2, 559-572(1901).

[2] Cramer, R. D. III; Patterson, D. E.; Bunce, J. D; "Comparative Molecular Field Analysis (CoMFA). Effect of Sahpe on Binding of Steroids to Carrier Proteins," *J. Am. Chem. Soc.* **110**, 5959-5967(1988).

[3] Wold, H.; "Nonlinear Estimation by Iterative Least Squares Procedures," *Festschrift for J. Neyman,* David, F. N., Ed, J. Wiley, NY (1966). Wold, H., Lyttkens, E.; "Nonlinear Iterative Partial Least Squares (NIPALS) Estimation Procedures," *Bull. Intern. Statist. Inst: Proc, 37th session, London* 1-15 (1969).

[4] Available from TRIPOS Associates Inc., 1699 S. Hanley Rd., St. Louis, MO

[5] Cramer, R. D. III; "BC(DEF) Parameters. 1. The Intrinsic Dimensionality of Intermolecular Interactions in the Liquid," *J. Am. Chem. Soc.* **102**,1837-1849(1980).

[6] Allen, M. S; Tan, Y; Trudell, M. Ml; Narayanan, K; Schindler, L. R; Martin, M. J; Schultz, C; Hagen, T. J; Koehler, K. F; Codding, P. W; Skolnick, P; Cook, J. M. "Synthetic and Computer-Assisted analyses

of the Pharmacophore for the Benzodiazepine Receptor Inverse Agonist Site," *J. Med. Chem*, **33**, 2343-2357(1990).

[7] Kim, K. H; Martin, Y; "Direct Prediction of Dissociation Constants (pKa's) of Clonidine-like Imidazoles, 2-substituted Imidazoles, and 1-Methyl-2-substituted-imidazoles from 3D structures Using a Comparative Molecular Field Analysis (CoMFA) Approach," *J. Med. Chem,* **34**, 2056-2060(1991).

[8] Kim, K. H; Martin, C. M; "Direct Prediction of Linear Free Energy Substituent Effects from 3D structures Using Comparative Molecular Field Analysis. 1. Electronic Effects of Substituted Benzoic Acids," *J. Org. Chem,* **56**, 2723-2729(1991).

[9] Good, A. C; Sung-Sau, S; Richards, W. G; "Structure-Activity Relationships from Molecular Similarity Matrices," *J. Med. Chem.* **36**, 433-438(1993).

[10] Good, A. C; Peterson, S. J; Richards, W. G; "QSAR's from Similarity Matrices. Technique Validation and Application in the Comparison of Different Similarity Evaluation Methods," *J. Med. Chem.* **36**, 2929-2937(1993).

[11] Jain, A. N; Koile, K; Chapman, D; "Compass: Predicting Biological Activities from Molecular Surface Properties. Performance Comparisons on a Steroid Benchmark," *J. Med. Chem,* **37**, 2315-2327(1994).

[12] Breslin, H. J; Kukla, M. J; Ludovici, D. W; Mohrbacher, R; Ho, W; Miranda, M; Rodgers, J. D; Hitchens, T. K; Leo, G; Gauthier, D. A; Ho, C. Y; Scott, M. K; De Clercq, E; Pauwels, R; Andries, K; Janssen, M. A. C; Janssen, P. A. J; "Synthesis and Anti-HIV-1 Activity of 4,5,6,7-Tetrahydro-5-methylimidazo- [4,5,1-jk][1,4]benzodiazepin-2(1H)-one (TIBO) Derivatives. 3," *J. Med. Chem.* **38**, 771-793(1995).

[13] Wagener, M., Sadowski, J. and Gasteiger, J., "Autocorrelation of Molecular Surface Properties for Modeling *Corticosteroid Binding Globulin* and Cytosolic *Ah* receptor Activity by Neural Networks," *J. Am. Chem. Soc.* **117**, 7769-7775(1995).

[14] Kubinyi, H., "A General View on Similarity and QSAR Studies," in *Computer-Assisted Lead Finding and Optimization, Current Tools for Medicinal Chemistry,* van de Waterbeemd, H., Testa, B. and Folkers, G. (Eds), Wiley-VCH (pub), Weinheim 1997, pp 7-28.

[15] Crippen, G. M., "Validation of EGSITE2, a Mixed Integer Program for Deducing Objective Site Models from Experimental Binding Data," *J. Med. Chem.* **40**, 3161-3172(1997).

[16] So, S. and Karplus, M., "Three-Dimensional Quantitative Structure-Activity Relationships from Molecular Similarity Matrices and Genetic Neural Networks, 1. Method and Validations," *J. Med. Chem.* **40**, 4347-4359(1997)

[17] Coats, E. A; "The CoMFA Steroids as a Benchmark Dataset for Development of 3D QSAR Methods, in 3D QSAR in Drug Design," in *3D QSAR in Drug Design, Vol III, Recent Advances* Ed. Kubinyi, H.; Folkers, G.; Martin, Y. C.; Klewar Academic Publishers, Dordrecht, the Netherlands, 1998, pp 199-213.

[18] Silverman, B. D. and D. E. Platt, "Comparative Molecular Moment Analysis (CoMMA): 3D-QSAR Without Molecular Alignment," In Journal of Medicinal Chemistry, vol. 39. 1995.

[19] Silverman, B. D., "The Thirty-one Benchmark Steroids Revisited: Comparative Molecular Moment Analysis (CoMMA) with Principal Component Regression," *Quant. Struct.-Act. Relat.*, **19**, 237-246(2000).

[20] Silverman, B. D. and D. E. Platt, "Comparative Molecular Moment Analysis (CoMMA): 3D-QSAR Without Molecular Alignment," In Journal of Medicinal Chemistry, vol. 39. 1995.

[21] Almoy, T., Haugland, E., "Calibration Methods for NIRS Instruments: A Theoretical Evaluation and Comparisons by Data Splitting and Simulations," *Applied Spectroscopy*, **48**, 327-332(1994).

[22] Faber, K., Kowalski, B. R.,"Propagation of Measurement Errors for the Validation of Predictions Obtained by Principal Component Regression and Partial Least Squares," *Journal of Chemometrics*, **11**, 181-238(1997)

[23] Lovely, C. J., Bhat, A. S., Coughenour, H. D., Gilbert, N. C., Brueggeimeier, R. W., " Synthesis and Biological Evaluation of 4-(Hydroxyalkyl)estradiols and Related Compounds" *J. Med. Chem.* **40**, 3756-3764(1997).

[24] Pavesi, A., "Relationships Between Transcriptional and Translational Control of Gene Expression in *Saccharomyces cerevisiae*: A Multiple Regression Analysis," *J. Mol. Evol.* (1999)48:133-141

[25] Cole, K. A., Krizman, D. B., Emmert-Buck, M. R., "The Genetics of Cancer – a 3D Model," *Genetics* (A *Nature* publication) **21**, 38-41(1999).

[26] Winzeler, E. A., Richards, D. R., Conway, A. R., Goldstein, A. L., Kalman, S., McCullough, M. J., McCusker, J. H., Stevens, D. A., Wodicka, L., Lockhart, D. J., Davis, R. W., "Direct Allelic Variation Scanning of the Yeast Genome," *Science,* **281**(5380), 1194-1197, Aug 21, 1998.

[27] Katritzky, A. R., Maran, U., Lobanov, V. S., Karelson, M., "Structurally Diverse Quantitative Sturcture-Property Relationship Correlations of Technologically Relevant Physical Properties," *J. Chem. Inf. Comput. Sci.*, **40**, 1-18(2000).

[28] Spector, S., Rosconi, M., Raleigh, D. P., "Conformational Analysis of Peptide Fragments Derived from the Peripheral Subunit-Binding Domain from the Pyruvate Dehydrogenase Multienzyme Complex of Bacil-

lus stearothermophilus: Evidence for Nonrandom Structure in the Unfolded State," *Biopolymers*, **49**, 29-40((1999).

[29] Hua, XY and Raleigh, D. "On the Global Architecture of Initiation Factor IF3: A Comparative Study of the Linker Regions from the Escherichia coli Protein and the Bacillus stearothermophilus Protein," *J. Mol. Biol.* **278**, 871-878(1998).

[30] Jolliffe, I. T.; *"Principal Component Analysis,"* Springer-Verlag, New York, 1986

[31] Press, W. H.; Teukolski, S. A.; Vetterling, W. T.; Flannery, B. P.; *"Numerical Recipes in C, 2nd ed,"* Cambridge University Press, NY, 1992.

[32] Freund, J. E.; *"Mathematical Statistics, 5th ed."* Prentice Hall, Upper Saddle River, NJ, 1992

[33] Bevington, P. R.; *"Data Reduction and Error Analysis for the Physical Sciences"*, McGraw-Hill, NY, 1969.

[34] Lott, W. F., "The optimal set of principal component restrictions on a least-squares regression," *Commun. Statist.*, **2** (1973), 449-464.

[35] Platt, D. E., L. Parida, Y. Gao, A. Floratos and I. Rigoutsos, "QSAR in Grossly Underdetermined Systems: Opportunities and Issues." In IBM Journal of Research and Development. In Press.

[36] Wold, S.; "Cross-Validatory Estimation of the Number of Components in Factor and Principlal Components Models," *Technometrics* **20**, 397405(1978).

[37] B. Efron, *"Jackknife, the Bootstrap & Other Resampling Plans"*, Society for Industrial and Applied Mathematics, 1982.

[38] Jackson, J. E., *"A User's Guide to Principal Components,"*, John Wiley and Sons, NY, 1991.

[39] Press, S. J.; *"Bayesian Statistics: Principles, Models and Applications"*, John Wiley and Sons, NY, 1989.

[40] Hill, R. C., Fomby, T. B. and Johnson, S. R. "Component Selection Normas for Principal Compoents Regression," *Commun. Statist.,* **A6**, 309-334(1977).

[41] Clark, M., Cramer, R. D. III, and van Opdenbosch, "Validation of the TRIPOS 5.2 Force Field," *J. Comput. Chem.*, **10**, 982-1012(1989).

[42] Frisch, M. J., Trucks, G. W., Head-Gordon, M., Gill, P. M. W., Wong, M. W., Foresman, J. B., Johnson, B. G., Schlegel, H. B., Robb, M. A., Replogle, E. S., Gomperts, R., Andres, J. L., Raghavachari, K., Binkley, J. S., Gonzalez, C., Martin, R. L., Fox, D. J., Defrees, D. J., Baker, J., Stewart, J. J. P., Pople, J. A., "Gaussian 92, Revision C," Gaussian Inc., 4415 Fifth Avenue, Pittsburgh, PA 15213.

Table 1: 21 CoMMA Descriptors

I_x amu-Å^2	I_y amu-Å^2	I_z amu-Å^2	p_x e-Å	p_y e-Å	p_z e-Å	p e-Å	q e-Å^2	q_{xx} e-Å^2	q_{yy} e-Å^2	d_x Å	d_y Å	d_z Å
931.63	4546.65	4938.82	0.4952	0.5140	0.1490	0.7291	11.0412	-5.6178	-2.3094	8.2273	5.4694	1.8269
790.16	4381.96	4861.82	0.3998	0.0961	0.2867	0.5013	14.9601	-5.3851	13.4133	11.6361	1.5378	13.7519
893.07	4578.96	4988.86	0.6979	0.0553	0.2499	0.7433	11.7459	-1.3790	11.2658	8.4244	2.9809	9.8699
622.02	2995.42	3316.53	0.7769	0.1862	0.1217	0.8081	3.2629	-0.0126	-2.9277	3.5541	2.3327	3.1487
544.40	2842.23	3196.95	0.1026	0.0197	0.2514	0.2723	8.9785	0.7687	-0.3653	4.6459	2.9329	2.4664
608.03	3247.75	3632.84	0.6221	0.0708	0.1236	0.6382	4.1172	-0.1809	0.2657	5.1398	5.7280	1.3064
536.55	2769.99	3123.35	0.4929	0.2050	0.2102	0.5737	6.0591	0.9080	3.7394	4.9107	5.1797	1.1889
842.76	2667.02	2772.00	0.4970	0.3198	0.5364	0.7982	7.5800	0.4016	3.7201	2.5517	2.3555	0.1187
698.47	3920.82	4309.91	0.1863	0.3089	0.1106	0.3773	11.4266	-5.9301	-3.4086	12.4445	6.4638	2.8828
761.81	4202.29	4531.81	0.2687	0.5392	0.1535	0.6217	13.8848	-8.4355	1.0865	5.6349	3.7610	2.1358
703.36	3773.53	4165.18	0.5814	0.2144	0.1383	0.6349	6.8832	-0.9245	1.4685	11.0534	5.3463	6.5989
631.35	3257.88	3583.00	0.2197	0.0098	0.3915	0.4490	7.7018	0.0950	0.0629	0.9324	2.6090	1.2987
782.59	4013.49	4335.64	0.3221	0.0227	0.1305	0.3483	3.1454	-0.4393	3.1271	24.7143	3.6345	1.3121
606.61	2954.31	3276.16	1.1555	0.1590	0.0024	1.1664	1.7074	-0.0166	-0.8155	3.2962	1.9849	1.8622
599.76	3103.51	3419.73	0.3864	0.1367	0.0271	0.4108	4.3961	-0.2734	-3.2590	2.2074	4.3425	6.4725
620.94	2861.25	3159.77	0.5992	0.3384	0.5415	0.8756	15.1479	-7.5869	5.2529	4.7018	4.0808	2.8027
688.81	2829.86	3072.09	0.6368	0.2115	0.4422	0.8036	5.6797	1.0759	1.0009	2.8061	3.1952	0.3680
882.22	4416.18	4928.04	0.5473	0.0291	0.4619	0.7168	17.9561	-7.4856	17.8694	8.0317	0.8073	9.3829
971.70	4604.94	5055.32	0.9058	0.2742	0.2166	0.9709	2.5270	-0.1950	1.6137	6.1636	3.4024	4.2125
956.56	4632.23	5113.40	0.8567	0.4816	0.2835	1.0228	13.2736	-3.4224	1.8752	4.9162	0.7754	3.7030
616.01	3002.07	3290.11	0.1640	0.3040	0.5811	0.6760	11.8327	-8.7428	8.6369	3.5667	0.5092	2.0745

Table 2: 21 CBG Activities

Activity
6.279
7.653
7.881
5.919
5.00
5.00
5.00
5.255
5.255
5.00
7.380
5.00
7.740
6.724
5.00
5.763
5.613
7.881
7.881
6.892
5.00

Table 3: CoMMA PCR Regression Coefficients

Coefficients	±	Errors
0.00152784	±	0.00261244
−0.00237925	±	0.00333417
0.00244299	±	0.00274996
−2.80199	±	3.10896
−2.07624	±	1.39958
0.366779	±	1.58565
6.12432	±	3.78857
−0.0112143	±	0.046603
0.0985031	±	0.0781397
−0.0269981	±	0.0384506
0.149761	±	0.0326788
0.00314725	±	0.0802196
0.167175	±	0.0555252
0.412245	±	1.16034

Table 4: CoMMA PCR Regression Coefficient Correlations

Correlation Coefficients													
1	-0.7434	0.6714	-0.3406	-0.3687	-0.5878	0.2629	0.0105	-0.1678	-0.0519	0.1478	0.2336	0.1432	-0.1786
-0.7434	1	-0.9930	0.5530	0.1990	0.4893	-0.4813	0.04850	-0.1776	0.07018	-0.4405	-0.2823	-0.3020	0.3947
0.6714	-0.9930	1	-0.5552	-0.1722	-0.4379	0.4855	-0.07479	0.2116	-0.07062	0.4391	0.2802	0.2941	-0.4289
-0.3406	0.5530	-0.5552	1	0.5665	0.6797	-0.9857	-0.2212	-0.6701	0.1622	-0.6445	-0.5786	-0.6073	0.7414
-0.3687	0.1990	-0.1722	0.5665	1	0.4534	-0.6146	-0.3651	-0.0976	0.4536	-0.2789	-0.4824	-0.2300	0.5763
-0.5878	0.4893	-0.4379	0.6797	0.4534	1	-0.6379	-0.3673	-0.3589	-0.2256	-0.2945	-0.3115	-0.1413	0.2345
0.2629	-0.4813	0.4855	-0.9857	-0.6146	-0.6379	1	0.2903	0.6870	-0.2255	0.6817	0.6005	0.6135	-0.7848
0.01046	0.0485	-0.0748	-0.2212	-0.3651	-0.3673	0.2903	1	0.4906	-0.2488	0.3298	0.08517	-0.0459	-0.2404
-0.1678	-0.1776	0.2116	-0.6701	-0.09764	-0.3589	0.6870	0.4906	1	-0.0062	0.5742	0.3093	0.3895	-0.5470
-0.0519	0.0702	-0.0706	0.1623	0.4536	-0.2256	-0.2255	-0.2488	-0.0062	1	-0.3953	-0.0030	-0.4772	0.4029
0.1478	-0.4405	0.4391	-0.6445	-0.2789	-0.2945	0.6817	0.3298	0.5742	-0.3953	1	0.2386	0.5582	-0.5498
0.2336	-0.2823	0.2802	-0.5786	-0.4824	-0.3115	0.6005	0.0852	0.3093	-0.0030	0.2386	1	0.3753	-0.6873
0.1432	-0.3020	0.2941	-0.6073	-0.2300	-0.1413	0.6135	-0.0459	0.3895	-0.4772	0.5582	0.3753	1	-0.5033
-0.1786	0.3947	-0.4289	0.7414	0.5763	0.2345	-0.7848	-0.2405	-0.5470	0.4029	-0.5498	-0.6873	-0.5033	1

Table 5: CoMMA PCR Regression Activity Prediction

Measured Activity	Predicted Activity	±	Errors
6.279	6.08868	±	0.240674
7.653	7.91285	±	0.266234
7.881	7.99297	±	0.288605
5.919	5.87558	±	0.22007
5.0	4.82527	±	0.28579
5.0	5.48766	±	0.232538
5.0	4.92669	±	0.317322
5.255	5.41805	±	0.329263
5.255	5.61347	±	0.248522
5.0	4.83734	±	0.307287
7.38	7.11644	±	0.253489
5.0	4.9226	±	0.28976
7.74	7.65107	±	0.346948
6.724	6.70262	±	0.330707
5.0	4.8961	±	0.295984
5.763	5.57878	±	0.308924
5.613	5.60417	±	0.203773
7.881	7.6102	±	0.311304
7.881	7.75545	±	0.300585
6.892	7.13346	±	0.315381
5.0	5.16656	±	0.32183

Table 6: One Component CoMMA PCR Regression Coefficients

Coefficients	$\pm$	Errors
0.00152784	$\pm$	$1.88815e-05$
-0.00237925	$\pm$	$2.94034e-05$
0.00244299	$\pm$	$3.01911e-05$
-2.80199	$\pm$	0.0346277
-2.07624	$\pm$	0.0256587
0.366779	$\pm$	0.00453274
6.12432	$\pm$	0.0756858
-0.0112143	$\pm$	0.000138589
0.0985031	$\pm$	0.00121733
-0.0269981	$\pm$	0.000333649
0.149761	$\pm$	0.00185078
0.00314725	$\pm$	$3.88944e-05$
0.167175	$\pm$	0.00206598
0.412245	$\pm$	0.00509462

Table 7: CoMMA One Component PCR Regression Coefficient Correlations

Correlation Coefficients													
1	-1	1	-1	-1	1	1	-1	1	-1	1	1	1	1
-1	1	-1	1	1	-1	-1	1	-1	1	-1	-1	-1	-1
1	-1	1	-1	-1	1	1	-1	1	-1	1	1	1	1
-1	1	-1	1	1	-1	-1	1	-1	1	-1	-1	-1	-1
-1	1	-1	1	1	-1	-1	1	-1	1	-1	-1	-1	-1
1	-1	1	-1	-1	1	1	-1	1	-1	1	1	1	1
1	-1	1	-1	-1	1	1	-1	1	-1	1	1	1	1
-1	1	-1	1	1	-1	-1	1	-1	1	-1	-1	-1	-1
1	-1	1	-1	-1	1	1	-1	1	-1	1	1	1	1
-1	1	-1	1	1	-1	-1	1	-1	1	-1	-1	-1	-1
1	-1	1	-1	-1	1	1	-1	1	-1	1	1	1	1
1	-1	1	-1	-1	1	1	-1	1	-1	1	1	1	1
1	-1	1	-1	-1	1	1	-1	1	-1	1	1	1	1
1	-1	1	-1	-1	1	1	-1	1	-1	1	1	1	1

Table 8: CoMMA One Component PCR Regression Activity Prediction

Measured Activity	Predicted Activity	±	Errors
6.279	6.08868	±	0.0752453
7.653	7.91285	±	0.0977889
7.881	7.99297	±	0.0987791
5.919	5.87558	±	0.0726118
5.0	4.82527	±	0.0596318
5.0	5.48766	±	0.0678179
5.0	4.92669	±	0.0608852
5.255	5.41805	±	0.0669575
5.255	5.61347	±	0.0693726
5.0	4.83734	±	0.0597811
7.38	7.11644	±	0.0879467
5.0	4.9226	±	0.0608347
7.74	7.65107	±	0.0945537
6.724	6.70262	±	0.0828326
5.0	4.8961	±	0.0605072
5.763	5.57878	±	0.068944
5.613	5.60417	±	0.0692577
7.881	7.6102	±	0.0940487
7.881	7.75545	±	0.0958438
6.892	7.13346	±	0.0881571
5.0	5.16656	±	0.0638496

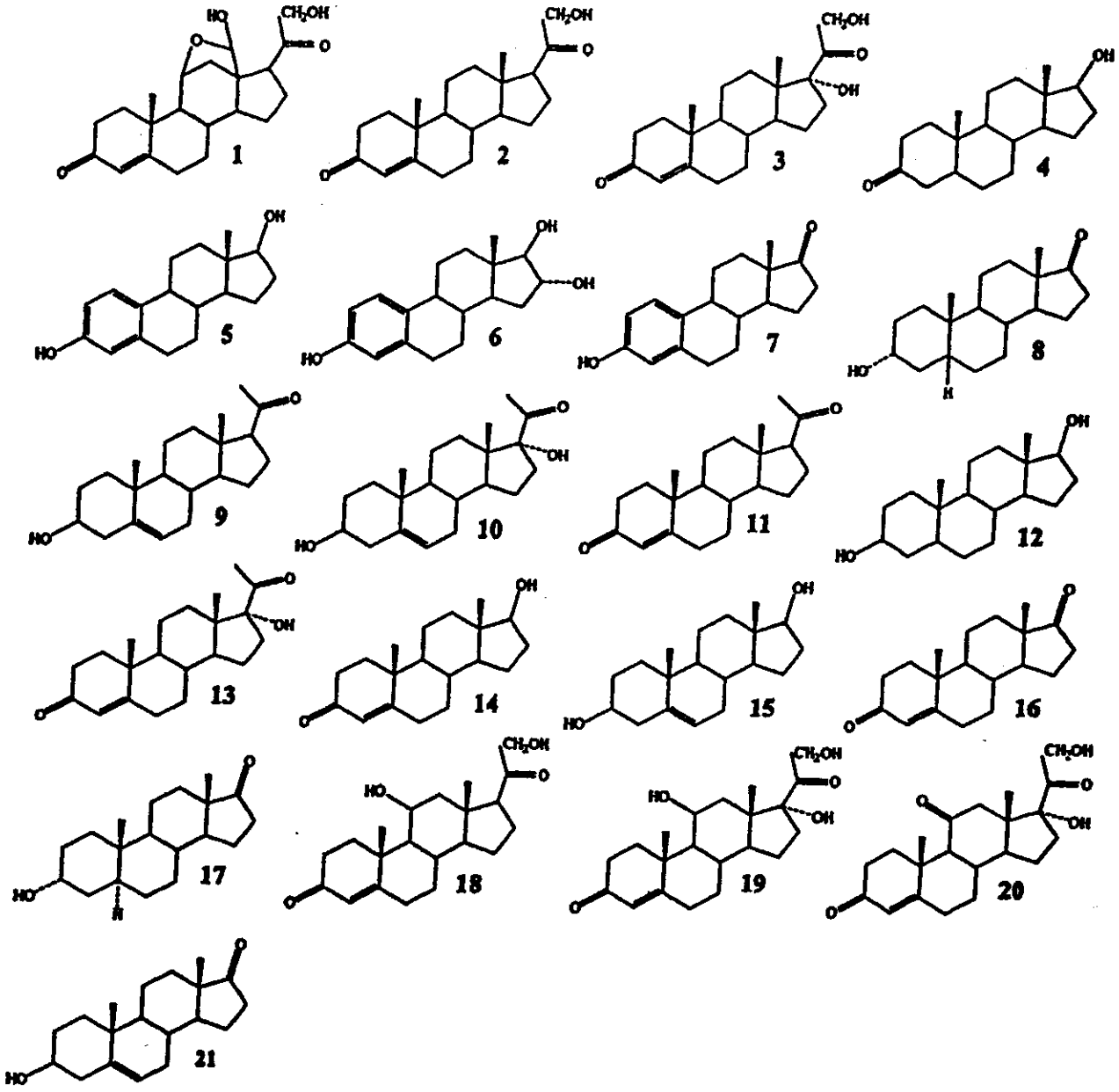

Figure 1: 21 corticosteroids with measured globulin binding activities.